Dual-modality Talking-metrics: 3D Visual-Audio Integrated Behaviometric Cues from Speakers

Dual-modality Talking-metrics: 3D Visual-Audio Integrated Behaviometric Cues from Speakers

Jie Zhang
1. Beihang University, Beijing, 100191, China
2. School of Informatics, University of Edinburgh
Email: zhangjie09@buaa.edu.cn

Korin Richmond
School of Informatics
University of Edinburgh
Email: korin@inf.ed.ac.uk

Robert B. Fisher
School of Informatics
University of Edinburgh
Email: rbf@inf.ed.ac.uk

***Abstract*—Face-based behaviometrics focus on dynamic biological signatures generated from face behaviors, which are informative and subject-specific for identity recognition. Most existing face behaviometrics rely on 2D visual features and thus are sensitive to pose or intensity variations. This paper presents a dual-modality behaviometrics algorithm (talking-metrics) that integrates 3D video and audio cues from a human face speaking a passphrase. Static and dynamic 3D face features are extracted algorithmically and audio features are transformed through a few learning models. We concatenate the top 18 discriminative 3D visual-audio features to represent the bi-modality and utilize an linear discrimant analysis (LDA) classifier for identity recognition. The experiments were conducted on a new publicly released dataset (S3DFM). Both qualitative feature distributions and quantitative comparison results show the feasibility of the proposed pipeline and the superiority over using each modality independently. A 98.5% cross-validation recognition rate over 60 subjects and 10 trials was achieved. An anti-spoofing test also demonstrates the robustness of the proposed method.**

I. Introduction

Face and speech biometrics are active topics in the fields of information security and forensics. The combination of dual biological modalities boosts biometric accuracy and especially increases spoofing resistance due to information compensation. In speech-related biometrics, the traits of speakers can be abstractly represented using either dynamic facial or audio features. 2D face biometrics have made great progress, with recognition rates approaching the level of human perception. However, a common inherent weakness with 2D face biometrics is sensitivity to head pose, scale, and intensity-related effects. Audio streams also suffer from interference from noisy backgrounds. In this paper, we combine 3D speech-related face metrics with audio metrics to give dual-modality "talking-metrics". The speech behavior creates a collaborative and text-guided facial modality. Compared with spontaneous facial expressions, speech has better repeatability and privacy, with the spoken text as a "passphrase". Additionally, 3D image information allows for real geometry-level biometrics, which are inherently robust against scale and pose. Compared with each individual modality, the dual-modality from the talking behavior is a promising approach for enhancing quantitative recognition accuracy and qualitative spoofing resistance.

A. Related Works

Visual-audio cooperative information has been exploited in applications including biometrics, speech recognition [1], person diarization [2], emotion recognition [3], and voice activity detection [4]. In the field of visual-audio dual-modality biometrics, the general pipelines follow three stages including raw feature extraction, feature fusion for joint representation, and identity decision. Modality fusion [5] can be categorized as feature-level (early) [6], [7], classifier-level (intermediate) [8], [9], [10], [11] or score/decision-level (late) [12], [13], [14] fusion. Existing visual-audio biometrics algorithms are listed in Table I. Note that all the algorithms used 2D intensity videos plus audio streams, while we focus on 3D shape videos.

Some joint visual-audio representations concatenate intensity-level visual features with audio features using a summation or maximum rule, while others train a deep fusion model, such as the cross-modal prediction model [10] or the joint Boltzmann machine model [11]. Quantitatively, some algorithms achieve nearly perfect results (higher than 95%) by using more audio or intensity information (more passwords, more face profiles, integrating teeth modality). We also achieve nearly perfect recognition rate (98.5%) and EER (1.33%) when only using the same password across 600 samples from 60 subjects. Qualitatively, the 3D face modality inherently benefits from less-sensitivity to the issues faced by the intensity modality, and it has been demonstrated that 3D face behaviors are even more informative and discriminative when describing a subject [15]. Thus, the dual-modality combining 3D facial dynamics and audio has more potential resistance to artificial spoofing, such as mask or voice play attacks. However, biometrics via 3D visual-audio modalities are a less-explored yet promising field.

In this paper, we propose a novel behaviometrics algorithm (talking-metrics) based on 3D visual-audio joint modalities generated from a 3D talking face.

Section II presents the proposed dual-modality behaviometrics that combine 3D talking face measurements with learned audio features for better accuracy and higher spoofing resistance, when compared to each modality as stand-alone. The pipeline consists of feature extraction from 3D talking face videos and synchronized audio streams, refined 3D visual-audio joint representation, feature training and identity classi-

TABLE I
DATA MODALITIES AND PERFORMANCE OF EXISTING VISUAL-AUDIO BIOMETRICS ALGORITHMS

Method/Year	Data modality	Subjects	Samples/Subject	Performance (%)
[12]/2007	indoor & outdoor face IM + audio	116	4 IM + 2 audio	TAR 97.5*; FAR 0.4*
[8]/2008	multi-profile face IM + multi-password audio	210	5	RR 96.7*
[6]/2009	face IV + audio	43	10	EER 5.9*
[7]/2010	open web TV shots + audio	10	40	RR 65.0*
[9]/2010	face IV + audio	43	10	RR 97.5
[13]/2010	face, teeth IM and audio	50	20	EER 1.6
[14]/2013	face IV + audio	88	12	RR 90.0*
[11]/2017	61 hours text-dependent & independent IV-audio	100	60	RR 98.0
Ours/2018	Speech-related 3D dynamic face video + audio	60	10	RR 98.5; EER 1.33

Note: IM-Intensity Image; IV-Intensity Video; RR-Recognition Rate. *Estimated from publication graphs.

fication.

Section II-A presents the first publicly available dataset about Speech-related 3D Face Motions - S3DFM Dataset[1]. The dataset consists of 600 2D-3D videos and audio sequences from 60 subjects, each with 10 repeatable samples.

Section III shows the experimental results conducted on the proposed S3DFM Dataset. The proposed pipeline achieves nearly perfect (98.5%) cross-validated recognition rates over 10 trials each by 60 subjects even with each speaking the same passphrase. It outperforms other biometrics pipelines based on individual 3D talking face, audio, and 2D visual-audio combined modalities.

II. PROPOSED METHOD

A. Overview

The proposed 3D visual-audio beheviometrics pipeline (as shown in Fig.1) has two main stages: (1) 3D visual-audio feature extraction; (2) unknown speaker recognition via pre-trained detectors. The visual-audio joint feature representation consists of static and dynamic 3D visual and audio feature extraction. A 3D talking face is represented using features constructed from the facial geometric structure and local shapes. The audio samples are originally represented as MFCCs. We train a generic Gaussian mixture based universal background model (GMM-UBM) and estimate a Total Variability matrix, which is then used for transforming each raw audio feature (MFCCs) into a higher-level feature in turn [16]. A subset of the concentrated visual-audio features are selected for use in another LDA classifier. More details follow below.

B. 3D Visual Representation

Given a 3D point cloud video and a pixel-wise registered 2D intensity video of a talking face, the 3D visual features are extracted from static (eyes and nose) and dynamic (mouth) regions of the 3D video, guided by the 2D intensity video. This is to avoid the influence from the noisy 3D video. The features are 3D distances between 3D facial landmarks (FLM) and principal curvatures (PC) of the 3D FLMs. They represent both static 3D facial geometry and talking-related 3D facial dynamics.

[1]http: //groups.inf.ed.ac.uk/trimbot2020/DYNAMICFACES

TABLE II
SPEECH-RELATED 3D DYNAMIC FACIAL PRIMITIVES.

Frame-level 3D Dynamic Facial Primitives	
Mouth width	DD_1
Mouth opening	DD_2
Max and min PCs of left mouth corner	DC_1, DC_2
Max and min PCs of right mouth corner	DC_3, DC_4
Max and min PCs of upper lip	DC_5, DC_6
Max and min PCs of lower lip	DC_7, DC_8
Frame-level 3D Static Facial Primitives	
Left eye width	SD_1
Right eye width	SD_2
L-R eye width	SD_3
Nose length	SD_4
Nose width	SD_5
Max and min PCs of nose bridge	SC_1, SC_2
Max and min PCs of nose tip	SC_3, SC_4

(PC: Principal Curvatures)

Specifically, the FLMs are the standard 68 points distributed around eyes, nose, mouth and cheek, extracted using [17]. We only selected 10 static FLMs from the eyes and nose, and 4 dynamic FLMs from the deforming mouth. The frame-level 3D facial features are extracted in 2 phases: leveraging an ensemble of regression trees for 2D FLM detection [17], then extracting corresponding 3D FLMs and neighborhoods from the pixel-wise registered 3D video. The derived 3D facial primitives from each frame are 5 Static FLM Distances $\mathbf{SD}_a = \{SD_a^t\}$, 2 Dynamic FLM Distances $\mathbf{DD}_b = \{DD_b^t\}$, 4 Static PCs $\mathbf{SC}_c = \{SC_c^t\}$, and 8 Dynamic PCs $\mathbf{DC}_d = \{DC_d^t\}$, as listed in Table II. Each facial primitive across a sequence generates a facial signature. Finally, we calculated statistics of each sequence signature and concentrated them into a holistic descriptor. Note that these statistics encode the dynamic properties of the talking face.

The i^{th} summary feature $f(\mathbf{x}_i)$ of either a static or a dynamic signature $\mathbf{x}_i = \{x_i^t\}$ is computed as

$$f(\mathbf{x}_i) = \begin{cases} \frac{1}{n}\sum\limits_t x_i^t & if \quad x_i^t = SD_a^t \ or \ SC_c^t \\ \begin{bmatrix} \max\limits_t(x_i^t) - \frac{1}{n}\sum\limits_t x_i^t \\ \frac{1}{n}\sum\limits_t x_i^t - \min\limits_t(x_i^t) \end{bmatrix} & if \quad x_i^t = DD_b^t \ or \ DC_d^t \end{cases} \tag{1}$$

By concatenating all the statistical properties of the facial

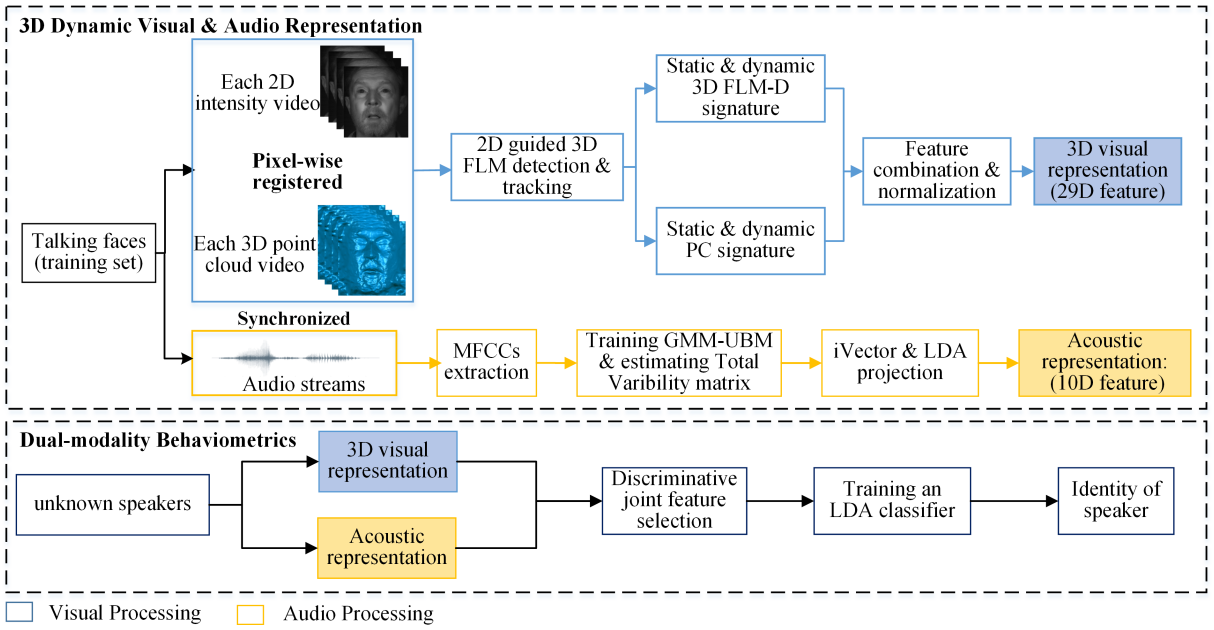

Fig. 1. The proposed 3D visual-audio behaviometrics pipeline. The top box details the 3D visual-audio joint representation, which is then used in the bottom box for dataset training and speaker testing.

signatures, a 3D talking face is finally represented as a 29D feature vector as

$$\mathbf{f} = \{w_a f(\mathbf{SD}_a), w_c f(\mathbf{SC}_c), w_{b'} f(\mathbf{DD}_b), w_{d'} f(\mathbf{DC}_d)\} \qquad (2)$$

where $w_a (a = 1 \cdots 5)$, $w_c (c = 1 \cdots 4)$, $w_{b'}$ ($b' = 1 \cdots 4$ is the number of statistical features of DD_b, so is twice the number of b), $w_{d'}$ ($d' = 1 \cdots 16$ is also twice the number of d) are all binary indexes that indicate whether the feature has been selected based on effectiveness for use (detailed in section II-D). The 3D video data and features have been previously used for person identification in [18].

C. Audio Representation

The process used to extract audio features is similar to that used in other recent work, for example [19]. It involves first calculating iVectors [16] and then applying linear discriminant analysis (LDA). Briefly, frame-level 42 dimensional spectral features were first extracted from each audio recording (down-sampled to 16kHz) in the form of mel-cepstra (12 dimensions, plus 0th order coefficient and energy, and their derived deltas and delta-deltas) using the Voicebox speech processing toolbox [20]. Then, for iVector processing, the training data was used to train a GMM-UBM with 256 mixture components, and a total variability matrix (T-matrix) for 40 dimensions. The training set iVectors and speaker labels were then used to derive an LDA projection to a reduced 10-dimensional space. The projection is to avoid the curse of dimensionality and thus improve the discriminativity of audio features. The UBM, T-matrix and LDA projection matrix thus obtained were used to process the test audio data in the same way.

D. Joint Feature Training and Classification

We combine the 29D visual features and 10D audio features into a 39D visual-audio joint feature vector. Since not all of the features are discriminative and informative enough for subject-specific identity classification, we select the first n features using a forward sequential feature selection strategy that minimizes the mean recognition error of an LDA classifier based on 5-fold cross validation. The selected top n discriminative features create a refined 3D visual-audio representation for the dual-modality talking-metrics.

For training and test, all the samples are separated into 3 splits (training, validation and test set). Each sample is represented by an n dimensional selected visual-audio feature vector. For each fold, an LDA classifier model is trained and tuned using the training and validation sets. In the test phase, a 3D visual-audio sample probe is input into the pre-trained LDA classifier for identity recognition. The first-rank class based on the nearest neighbor distance will be the identity of the test person.

III. Results and Discussions

To the best of our knowledge, there is no existing publicly released dataset on talking-related 3D dynamic face videos and audio streams. We construct Edinburgh Speech-related 3D Facial Motion Dataset (S3DFM), which enables research on joint 3D visual-audio recognition. The experiments were conducted using the proposed S3DFM dataset to investigate the performance and robustness to spoofing of the proposed method.

A. S3DFM Dataset

The new dataset contains 600 samples acquired from 60 subjects (each subject has 10 samples), where there are 37

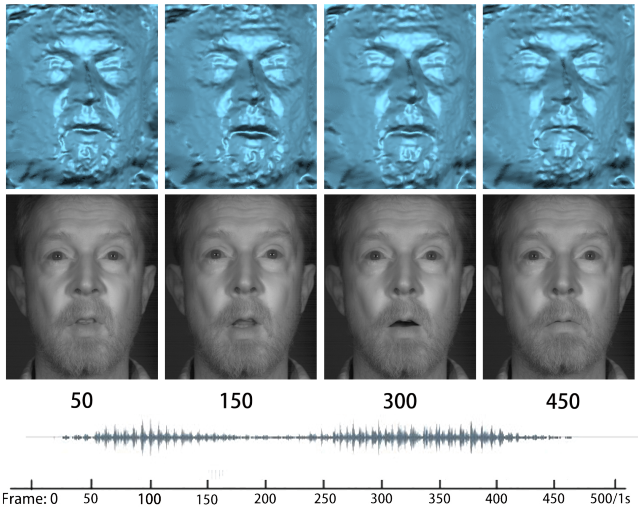

Fig. 2. Example data: top: 4 representative 3D point cloud frames; mid: 4 pixel-wise registered 2D infrared intensity frames; bottom: audio sample.

males and 23 females. The subjects cover more than 15 nationalities, and their ages range from 16 to 73 years old. In data acquisition, each subject was asked to spontaneously repeat a short passphrase - 'Nihao' 10 times in front of a binocular stereo video sensor [21] and a microphone. Each sample consists of a 500 fps 3D point cloud video with the resolution of 600 $\times$ 600 points, a pixel-wise registered 2D infrared intensity video with the resolution of 600 $\times$ 600 pixels and a synchronized audio stream with the sampling frequency of 16kHz. To increase the difficulty, we used the same passphrase across all the subjects. An example sample from a subject is shown in Fig. 2. All the experiments below used 5-fold cross-validation. The training, validation, and test set are respectively with 6 samples, 2 samples, and 2 samples each from the 60 subjects.

B. Feature Distribution

To qualitatively evaluate the separability of features, we projected (using PCA) 4 sets of features into a 2D space respectively and compare their distributions. The investigated feature sets are selected 3D visual-audio features (Selected 3D VA), 3D static visual features (3D SV), 3D dynamic visual features (3D DV), and audio features. The distributions of the 4 feature sets from one fold are shown in Fig.3. The 2 test samples from one class are linked with the center of the remaining 8 samples from the same class using a line to show the intra-separability of the class.

Overall, all the distributions of the 4 feature sets show some separability of classes, although the separability of the audio samples is not large compared with its intra-separability. The 3D SV samples have both small inter and intra-separabilities. The inter-separability of the 3D DV samples is better, while a few samples are not clustered well. The selected joint visual-audio features benefit from the 3 individual feature types, although there are still a few outlier samples that lie relatively far from the center of the corresponding class.

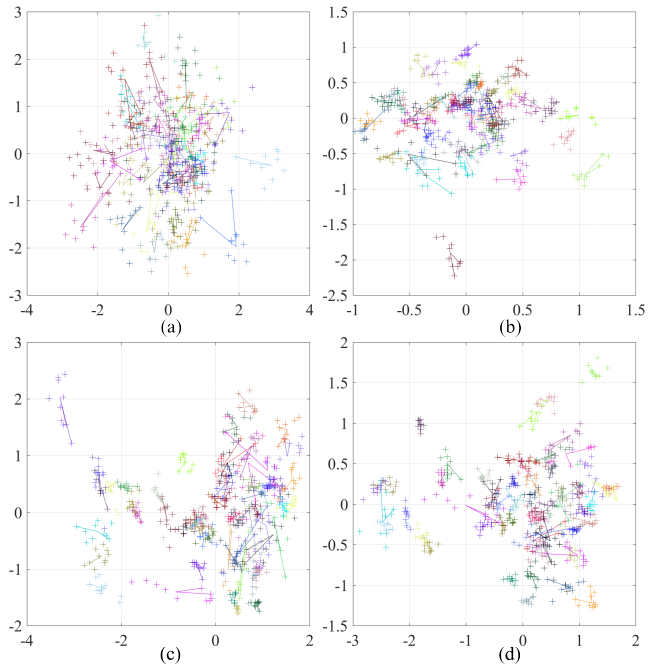

Fig. 3. The distribution of feature sets: (a) audio features; (b) 3D dynamic visual (3D DV) features; (c) 3D static visual (3D SV) features; (d) selected 3D visual-audio features.

C. Feature Selection

The most effective features are selected from the full 39D visual-audio joint features to maximize the discriminativity power and to remove noisy components. To compare the contribution of each modality's features individually, we did independent feature selection on 3D visual features (3D SV + 3D DV) and audio features. Fig.4 shows feature selection error for pure 3D visual features, pure audio features, and 3D joint visual-audio features, respectively, where the number of original 3D visual features is 29 and the number of audio features is 10. The iterative optimization of the sequential selection algorithm minimizes the mean recognition error (%) achieved by an LDA classifier over the 600 samples. The selection was performed using 5-fold cross-validation over the full dataset.

One can see that (1) the pure 3D visual feature selection shows that the first 13 visual features achieve the best performance with a recognition rate of 95.0%, while the main improvement is given by the first 7 visual features, with a recognition rate of 93.7%. The detailed feature numbers are listed in Table III. (2) The pure audio feature selection shows that all of the 10 audio features are useful for recognition, with the best recognition rate of 94.7%. (3) The feature selection over the full visual-audio joint features shows that the first 18 visual-audio joint features are the most useful for our task, with the overall best recognition rate of 98.5%. The order of the sequentially selected 18 features is listed in Table IV. The main improvement is given by the first 10 joint features, with the minimum recognition rate of 96.7%.

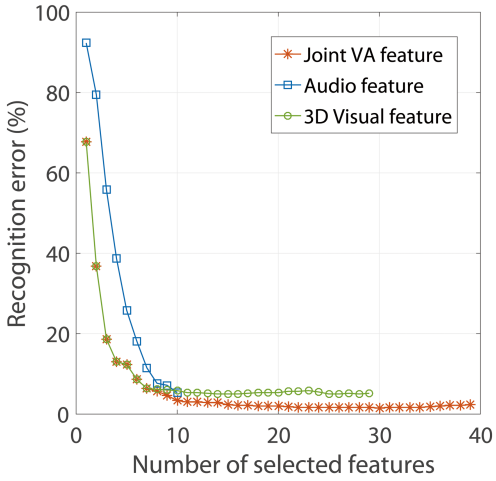

Fig. 4. Recognition error for different numbers of selected features.

TABLE III
FEATURE NUMBERS BEFORE AND AFTER FORWARD SEQUENTIAL SELECTION

Original Feature Set	Original Feature Number	Selected Number
3D V (S + D)	29	13
Audio	10	10
Joint 3D visual-audio	39	18

TABLE IV
SEQUENTIALLY SELECTED VISUAL-AUDIO JOINT FEATURES

Order	Type	Feature	Order	Type	Feature
1	3D SV	$f(SD_3)$	10	Audio	Audio 7
2	3D SV	$f(SD_5)$	11	3D DV	$f(DC_3)_2$
3	3D SV	$f(SD_2)$	12	Audio	Audio 10
4	3D SV	$f(SD_1)$	13	Audio	Audio 1
5	3D DV	$f(DD_2)_1$	14	Audio	Audio 2
6	3D SV	$f(SD_4)$	15	3D DV	$f(DD_1)_1$
7	3D DV	$f(DD_2)_2$	16	3D SV	$f(SC_2)$
8	Audio	Audio 3	17	3D DV	$f(DC_5)_2$
9	Audio	Audio 6	18	3D DV	$f(DC_6)_1$

(The numbers of the audio features do not have explicit interpretations due to the LDA refinement.)

D. Performance of 2D and 3D Feature Combinations

We investigate the performance of selected 3D visual-audio joint features for person identification, in comparison to several other feature combination pipelines based on 3D SV and/or 3D DV and/or audio and/or algorithmic 2D visual (2D V) features, and a state-of-the-art "deep" face descriptor (FaceNet) [22]. The 2D features were extracted from a 2D intensity set with 2400 images (60 subjects×40 images/subject) randomly selected from the proposed S3DFM Dataset. The algorithmic 2D facial features are Gabor Magnitude (GM) [23] and Phase Congruency (PhaseC) [24] for comparison. The deep-net-based pipeline was trained via fine-tuning a pre-trained FaceNet model. For the pipeline based on algorithmic 2D V, we trained an LDA classifier and tuned its projection parameter over the training and validation sets using 5-fold cross-validation. The parameters with the minimum mean recognition error were used in the test phase. The comparison results are shown in Fig. 5 using Cumulative Match Characteristic (CMC) curves. The CMC curves show that the accuracy of the algorithm varies with 'Top N' correctness, for different values of N. A decision is correct if the true identity is in the top N scores.

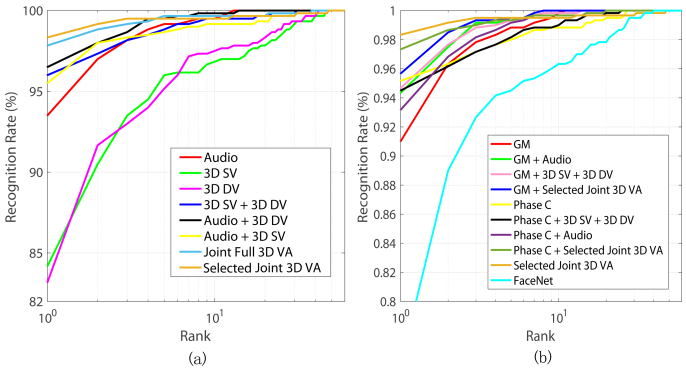

Fig. 5. Comparison of retrieval results of varying pipelines: (a) 3D +/ audio features; (b) 2D +/ 3D +/ audio features. (The results in the 2 sub-figures can be compared together, while are separated just for easier viewing)

The FaceNet-based pipeline show a overall mean recognition rate of 76.5% over the 2400 intensity images. As the FaceNet-based descriptors have 4096 dimensions and the size of the dataset is limited, we do not think that the FaceNet is the most compatible solution with our data. The results in Fig. 5 show that the pipeline based on the selected 3D visual-audio features outperforms all of the other pipelines, with the highest first-rank recognition rate of 98.5%, followed by the full 3D visual-audio features with 97.9% and the feature combination of Phase Congruency 2D visual features and the selected 3D visual-audio features, with 97.3%. In Fig.5a, it is obvious that the 3 individual feature sets (3D SV, 3D DV, audio) performed relatively poorly. Therefore, each individual modality benefits from the other modalities by information compensation.

E. Biometric Verification with Spoofing Attacks

Spoofing attacks are one of the greatest challenges in biometrics. To simulate audio and face spoofing attacks, we separate all the test samples into three classes, including 60 genuine clients, 30 3D V (3D DV + 3D SV) spoofing attacks, and 30 audio spoofing attacks.

For the 3D V spoofing attacks, the 3D V features of person B were replaced by the 3D V features from target subject A (as if person B with their own audio features was pretending to be person A). Similarly, for audio spoofing attacks, the audio features of person B were replaced by the audio features from target person A (as if person B with their own video features was pretending to be person A.). The test was conducted using 5-fold cross validation. The distributions of the genuine scores, 3D V impostor scores, and audio impostor scores are shown in Fig.6a, and the ROC curve of the closed-set verification is shown in Fig.6b.

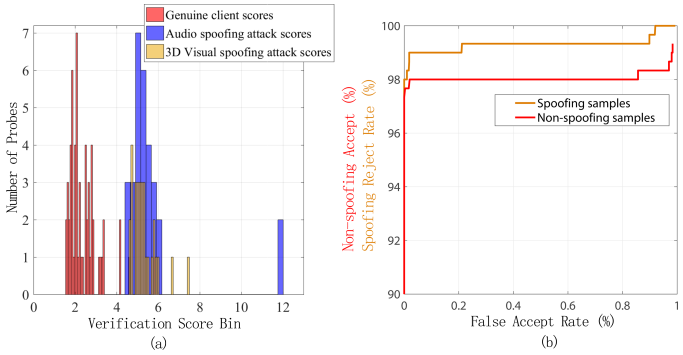

Fig. 6. (a) Distributions of the three sets of samples with/without spoofing attacks; (b) ROC curves for accepting true probes and rejecting spoofing attacks, obtained by varying the verification score threshold.

It is obvious that the genuine scores are separable from the impostor scores (audio + 3D V spoofing attacks), which demonstrates the discriminative power of the individual modality features and the feasibility of the 3D visual-audio joint feature representation for anti-spoofing. The ROC gives an EER of 1.33%. Overall, the proposed 3D visual-audio joint pipeline is robust against the audio and 3D V spoofing attacks.

IV. CONCLUSIONS

This paper presents a novel dual-modality behaviometrics approach (talking-metrics) based on a 3D visual-audio joint feature representation. Also, we presented a novel 2D-3D facial video plus audio dataset concerning talking-related behavior. The features of each modality are extracted or learned separately, and are integrated into a 18D summary feature vector by sequentially selecting the top discriminative features. The proposed dual-modality behaviometrics achieves the best performance over the biometric identification and anti-spoofing verification tests on our released dataset (S3DFM). Our works explore a novel discriminative combination of talking-related 3D dynamic face and audio information and thus provide a new biometric approach. The discriminability of the 3D face behavior and mutual compensation between the dual modalities improve the level of biometric security.

We note that this excellent performance is achieved even with all subjects speaking the same passphrase. Future work could explore the benefits of each subject using a unique passphrase. Another possible future direction could explore performance when the facial poses are changing unpredictably while speaking, or deep-net-based feature extraction.

ACKNOWLEDGMENTS

The work was supported by the funding from the China Scholarship Council (CSC) under grant 201606020087. We would like to thank all the participants for their contribution to the S3DFM Database.

REFERENCES

[1] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning dynamic stream weights for coupled-hmm-based audio-visual speech recognition," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 23, no. 5, pp. 863–876, 2015.

[2] G. Paul, K. Elie, M. Sylvain, O. Jean-Marc, and D. Paul, "A conditional random field approach for audio-visual people diarization," in *2014 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 116–120.

[3] F. Noroozi, M. Marjanovic, A. Njegus, and S. Escalera, "Audio-visual emotion recognition in video clips," *IEEE Trans. Affective Computing*, vol. PP, no. 99, pp. 1949–3045, June 2017.

[4] D. Dov, R. Talmon, and I. Cohen, "Audio-visual voice activity detection using diffusion maps," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 23, no. 4, pp. 732–745, 2015.

[5] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010.

[6] D. Shah, K. J. Han, and S. S. Narayanan, "A low-complexity dynamic face-voice feature fusion approach to multimodal person recognition," in *2009 11th IEEE Int. Symp. Multimedia*, Dec 2009, pp. 24–31.

[7] R. M. Jiang, A. H. Sadka, and D. Crookes, "Multimodal biometric human recognition for perceptual human-computer interaction," *IEEE Trans. Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 676–681, Nov 2010.

[8] A. Das, O. K. Manyam, and M. Tapaswi, "Multi-feature audio-visual person recognition," in *2008 IEEE Workshop on Machine Learning for Signal Processing*, Oct 2008, pp. 227–232.

[9] H. Zheng, M. Wang, and Z. Li, "Audio-visual speaker identification with multi-view distance metric learning," in *2010 IEEE Int. Conf. Image Processing*, Sept 2010, pp. 4561–4564.

[10] S. Petridis and M. Pantic, "Prediction-based audiovisual fusion for classification of non-linguistic vocalisations," *IEEE Trans. Affective Computing*, vol. 7, no. 1, pp. 45–58, 2016.

[11] M. R. Alam, M. Bennamoun, R. Togneri, and F. A. Sohel, "A joint deep boltzmann machine (jdbm) model for person identification using mobile phone data," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 317–326, 2017.

[12] H. Vajaria, T. Islam, P. K. Mohanty, S. Sarkar, R. Sankar, and R. Kasturi, "Evaluation and analysis of a face and voice outdoor multi-biometric system," *Patt. Recog. Let.*, vol. 28, no. 12, pp. 1572 – 1580, 2007.

[13] D. J. Kim, K. W. Chung, and K. S. Hong, "Person authentication using face, teeth and voice modalities for mobile device security," *IEEE Trans. Consumer Electronics*, vol. 56, no. 4, pp. 2678–2685, November 2010.

[14] M. R. Alam, R. Togneri, F. Sohel, M. Bennamoun, and I. Naseem, "Linear regression-based classifier for audio visual person identification," in *2013 1st Int. Conf. Communications, Signal Processing, and their Applications (ICCSPA)*, Feb 2013, pp. 1–5.

[15] L. Benedikt, V. Kajic, D. Cosker, P. L. Rosin, and A. D. Marshall, "Assessing the uniqueness and permanence of facial actions for use in biometric applications," *IEEE Trans. Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 3, pp. 449–460, 2010.

[16] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[17] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[18] J. Zhang and R. B. Fisher, "Visual passphrase: Behaviometrics via speech-related 3d facial dynamics," *Under review*, 2018.

[19] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, and M. Todisco, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, June 2017.

[20] "Voicebox: speech processing toolbox for matlab," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html Jan.12, 2018.

[21] "Dimensional imaging (di4d)," http://www.di4d.com/ Jan.12, 2018.

[22] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 815–823.

[23] V. Štruc and N. Pavešić, "The complete gabor-fisher classifier for robust face recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, p. 31, 2010.

[24] S. Gundimada, V. K. Asari, and N. Gudur, "Face recognition in multi-sensor images based on a novel modular feature selection technique," *Information Fusion*, vol. 11, no. 2, pp. 124–132, 2010.